HE SAID, SHE SAID

“What do you do?” She asked,

Computers, I said.

“Oh that’s interesting, what kind of computers?”

Fast computers, I muttered.

Still interested (barely), “what type of fast computers?”

Very fast computers.

She moved to another seat at the bar.

If she had been at all persistent, she would have learned that I work on:

> *Low cost scalable very fast computers that you can assemble in one day and install all the software you need to achieve a GigaFlops performance at $50K and that they are called "Beowulf"*

Her loss.

Pile of PCs
Unthinkable, Obvious, Inevitable
A Beowulf Perspective

Thomas Sterling

NASA Jet Propulsion Laboratory
&
California Institute of Technology

April 10, 1997

ELCO COMPUTERS

229 S. Raymond Ave.
Alhambra, CA 91801
Telephone *818) 284-3281
FAX (818) 284-4871
SVC (818) 284-7018

Calif Institute of Technol
ATTN: Aero #52
391 S. Holliston Ave.
Pasadena, CA 91125
(818) 356-6811

INVOICE NO. SO - 339458
INVOICE DATE: 09/18/96

SHIP Date	Request Date 9/18/96	Ship Via PU	Freight Shipping Point
Salesman Sunny UY	Customer P.O. Quotation		Payment Terms Net 20

ITEM DESCRIPTION	ORDER	SHIP	B/O	PRICE	EXTENDED
700-MINMAW 3700ATX FOR PENTIUM PRO, MID TOWER CASE W/200W UL	17			99.00	1,683.00
Enp-Intel Venus Pentium Pro	17			300.00	5,100.00
o-Intel CPU Pentium Pro 200	17			650.00	10,625.00
Pentium Pro Cooling Fan	17			0.00	0.00
.52-WSTRN Dgtl. 32500 2.5GB IDE	36			295.00	10,620.00
44P-TEAC 1.44MB White (PS2)	2			23.00	46.00
D-R-ACER 685A/787 8x CD-ROM	2			85.00	170.00
014 Fujitsu 4720 101 Key	2			37.00	74.00
PCI-D-Link DFE-500 TX 100 Base T	17			85.00	1,445.00
BCD-4 to 1 ABCD 5H15 VGA Monitor/KBSwitch Box	5			25.00	125.00
Extension/6’ Keyboard Cable	17			5.00	85.00
GA - Extension/ Monitor Cable	17			6.00	102.00

 Continued on Next Page

ELCO COMPUTERS

229 S. Raymond Ave.
Alhambra, CA 91801
Telephone *818) 284-3281
FAX (818) 284-4871
SVC (818) 284-7018

Calif Institute of Technol
ATTN: Aero #52
391 S. Holliston Ave.
Pasadena, CA 91125
(818) 356-6811

INVOICE NO. SO - 339458
INVOICE DATE: 09/18/96

SHIP Date	Request Date 9/18/96	Ship Via PU	Freight Shipping Point
Salesman Sunny UY	Customer P.O. Quotation		Payment Terms Net 20

ITEM DESCRIPTION	ORDER	SHIP	B/O	PRICE	EXTENDED
Cabethp-Patch Calbe 14' LVL 5 TW/Pair ICC	17			12.00	204.00
Simede--EDP 8x32 32 MB 6ons	68			168.00	11,424.00
TR11--Generic VGA-16 1024 x 768 W/512K Trident 9000 Chipset	17			23.00	391.20
NOR14GE - Northstar Gem14NI 14" SVGA Non-interlaced (E to E)	1			208.00	208.00
NEC21XP-NEC XP21 21" Multisync 1600 x 1200, color	1			1,899.00	1,899.00
GAATIC-ATI-GXULT Win Turbo PCI-64	1			384.00	384.00
SBAWE32-SB AWE32 PnP IDE W/ASP	1			210.00	210.00
SPEYAMY-Yamaha YSTM15 Speaker	1			75.00	75.00
MICSERO-Microsoft OEM	1			24.00	24.00

Sub-total Amount	$44,894.00
Sales Tax	3,703.76
INVOICE TOTAL	$48,597.76

MAJOR IDEAS

- Something special has happened
- A new business model for high performance computing
- Challenges for near future

Performance

Warren-Salmon comparative results:

Site	Platform	Nodes	Time(s)	GigaFLOPS	MegaFLOPS/node
LANL	TMC CM-5	512	140.7	14.06	27.5
Cal Tech	Intel Paragon	512	144.4	13.7	26.8
NRL	TMC CM-5E	256	171	11.57	45.2
Cal Tech	Intel Data	512	199.3	10.02	19.6
NAS	IBM SP-2	128	281.9	9.52	74.4
JPL	Cray T3D	256	338	7.94	31
LANL	TMC CM-5 (no VU)	256	754.6	2.62	5.1
SC’96	Beowulf: Loki+Hyglac	32	1218	2.19	68.4

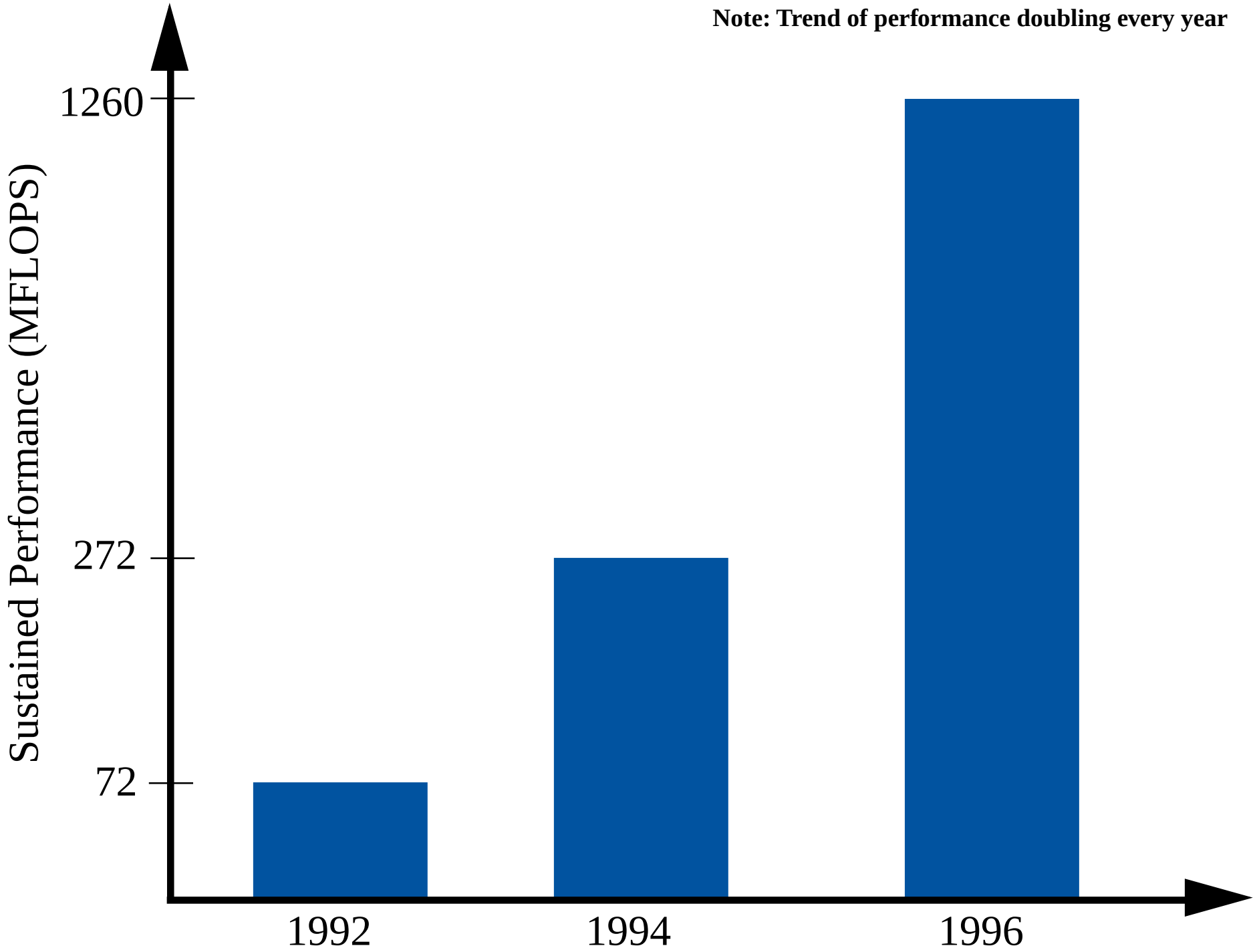
Incremental Advances in a Nonlinear Tradeoff Space:

Punctuated Equilibrium in Clustered Computing

- Drastic reduction in vendor supported HPC
- Technology of mass market COTS converges with workstations
- PC hosted software environments achieve workstation sophistication
- Network hardware and software enable balanced clusters
- MPPs establish low level of expectation

EXOTHERMIC PARALLEL COMPUTING

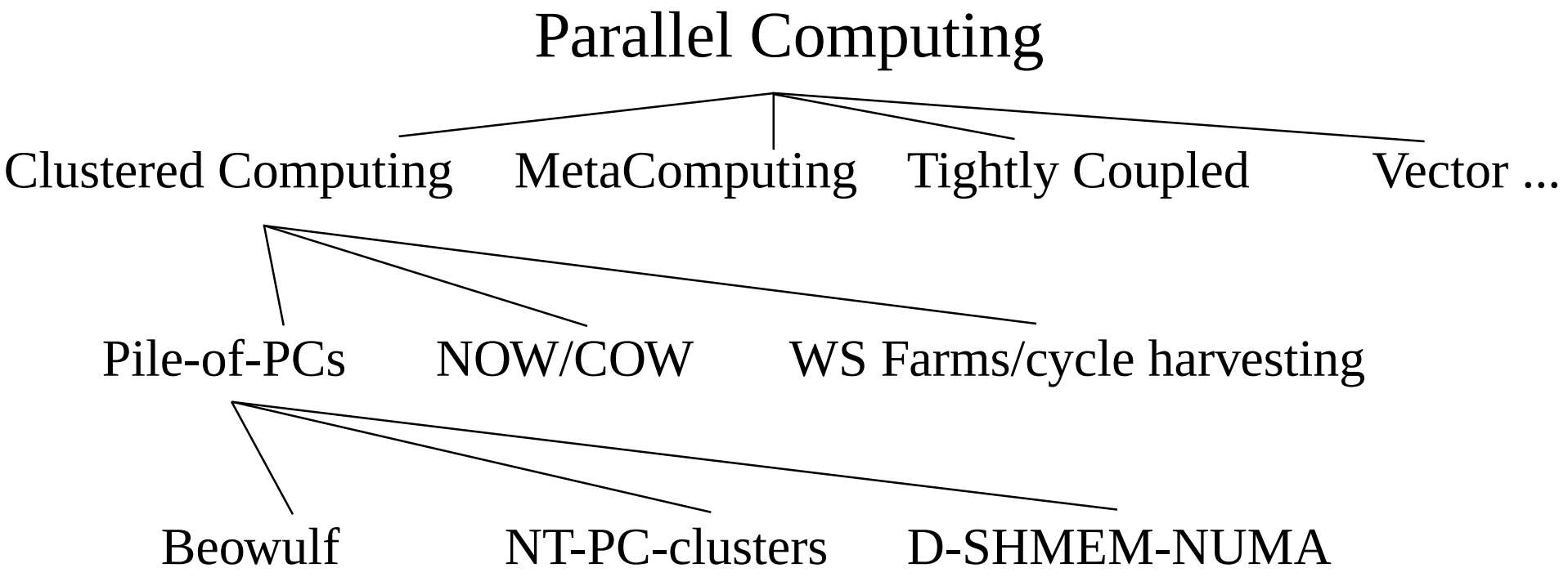
- What's the bid deal - except everything has changed
- Big Iron versus Big Computing
  - It's not Supercomputing if it's easy
  - It's not Supercomputing if it's cheap
- Parallel processing without parallel computing vendors
- Just-in-place configuration
- In phase with the Technology Wave
- Anyone (almost) can/is doing it: everything's available, no magic
- No artificial price structure, you pay what you get for - not more


Note: Trend of performance doubling every year
Sustained Performance (MFLOPS)
1260
272
72
1992
1994
1996

TAXONOMY

According to Tron

(incomplete, inconsistent, and entirely biased)



BEOWULF

Harnessing the Power of Parallelism at the price of PCs
Parts at Dawn, Processing by Dusk

- Mass Market COTS
  - good performance, best price-performance
  - multiple indistinguishable distributors and vendors
- Open Unix-like O/S with source code
  - e.g. Linux, BSD
- Distributed memory, message passing programming model
  - PVM, MPI
- User Application Driven
  - components
  - topologies
- Application Domains
  - single user, data heavey terminals
  - mass storage
  - scientific computation
  - emulation, simulation
  - education

BEOWULF

A Short History (appropriately distorted)

- Big bang, Galaxy formation, Earth cools, Dinosaurs go the way of supercomputers, Homo Erectus lives in caves for a million years waiting for ISDN, Gates invents vaporware - IBM sues for rights, Intel finally builds a real computer, The Great Extinction: Encore, TMC, BBN, KSR,
- Alliant, ETA, CCC, CRI, Convex, ...
- Meanwhile, a cold and lonely grad student in Finland ...
- A grad student wanabe working for NSA writes ethernet drivers, BUT can’t send messages out, so (EUREKA!) sends it to himself
- NASA needs a few good cycles
- Genesis of Beowulf - combines low cost x 86, Ethernet, Linus, clustered architecture, message passing programming
- Beowulf adds - ethernet drivers, channel bonding, advanced topologies, applications, ensemble management tools
- Three generations later: 1.25 GFLOPS sustained performance/$50K demonstrated at Supercomputing ‘96

Linux

- Robust Unix-like operating system
- Low initial cost
- Readily available source code
- Royalty-free redistribution
- Distributions provide ready vehicle for technology transfer

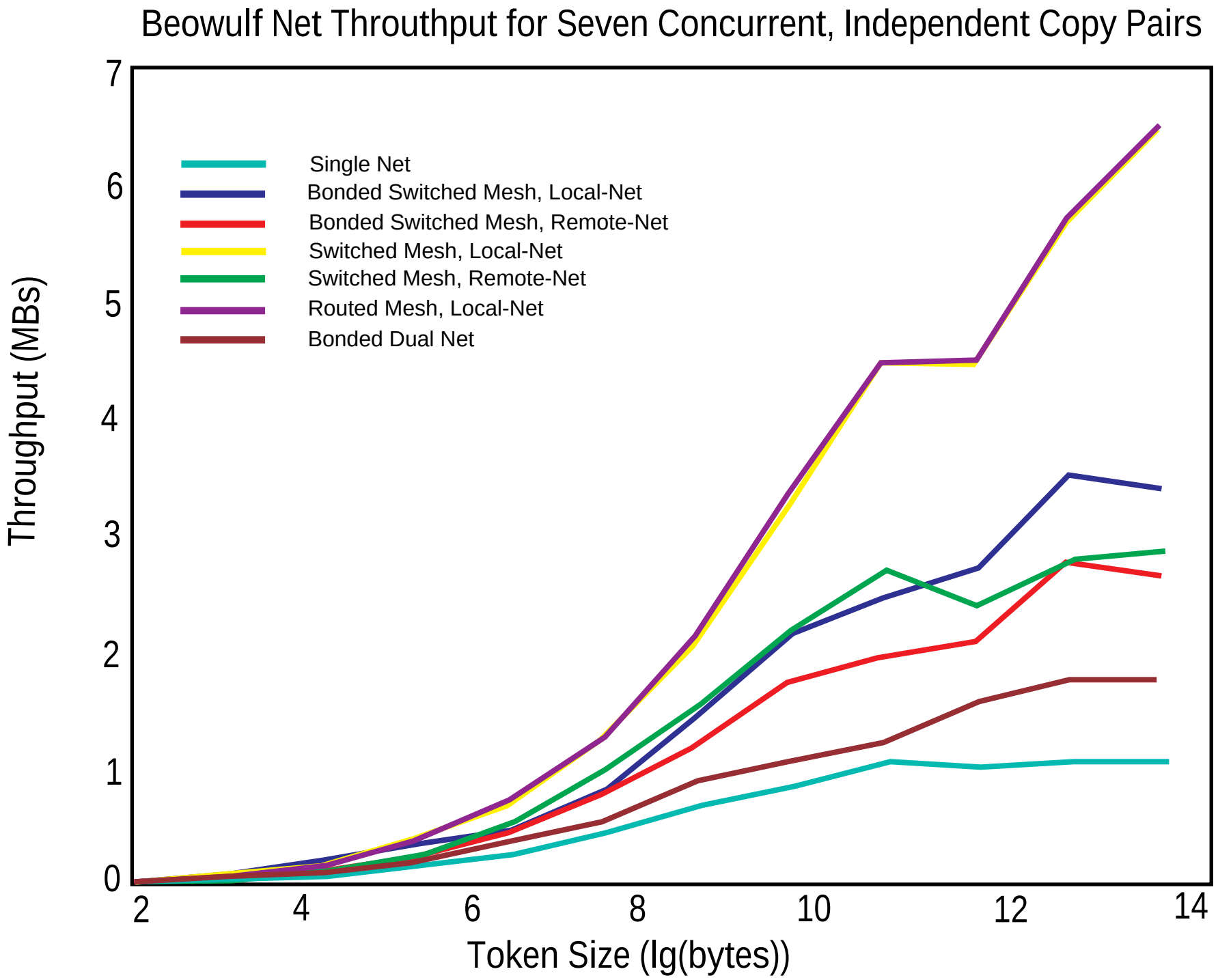
NOW WAS THEN

- Clusters of workstations may be ephemeral
  (but so may be workstations themselves)
- HP to use Intel parts; DEC targeting Alpha to PCs
- Restricted and closed operating systems
- Performance is comparable to PCs
- Markup heavy
- “Vendor Envy”

CHALLENGES

An applied research agenda for the Pile-of-PC community

- Scalability
- managing ensembles
- generality
- distributed shared memory?
- Robustness & system administration
- longevity of investment


Beowulf Net Throughput for Seven Concurrent, Independent Copy Pairs
Throughput (MBs)
Token Size (lg(bytes))
0
1
2
3
4
5
6
7
2
4
6
8
10
12
14
Single Net
Bonded Switched Mesh, Local-Net
Bonded Switched Mesh, Remote-Net
Switched Mesh, Local-Net
Switched Mesh, Remote-Net
Routed Mesh, Local-Net
Bonded Dual Net

SCALABILITY

- Network Technology
- Network Topology
- Network Cost
- Bandwidth versus Latency

MANAGING ENSEMBLES

- Global Name Spaces
- Virtualizing Process Ids
- Parallel versions of basic commands
- Parallel I/O
- Beotools, Beoworks, Beoware, Grendel
- Advanced Programming Models; BSP, distributed OOP

GENERALITY

- Algorithm Parallelism, granularity
- Latency tolerant algorithms
- Reduced communication latency
- Automatic load balancing

LONGEVITY OF INVESTMENT

- Technology moving very fast
- Mix of multiple generations
- Different types of performance
- Heterogeneous computing

CONCLUSIONS

- Unthinkable - but doable
- Obvious and needed
- Inevitable - unstoppable
- Role of vendors is to provide reliable building blocks